Biostatistics Primer

What a Clinician Ought to Know: Subgroup Analyses

Helen Barraclough, MSc, and Ramaswamy Govindan, MD†‡*

Abstract: Large randomized phase III prospective studies continue to redefine the standard of therapy in medical practice. Often when studies do not meet the primary endpoint, it is common to explore possible benefits in specific subgroups of patients. In addition, these analyses may also be done, even in the case of a positive trial to find subsets of patients where the therapy is especially effective or ineffective. These unplanned subgroup analyses are justified to maximize the information that can be obtained from a study and to generate new hypotheses. Unfortunately, however, they are too often overinterpreted or misused in the hope of resurrecting a failed study. It is important to distinguish these overinterpreted, misused, and unplanned subgroup analyses from those prespecified and well-designed subgroup analyses. This overview provides a practical guide to the interpretation of subgroup analyses.

Key Words: Biostatistics, Subgroup analysis.

(*J Thorac Oncol.* 2010;5: 741–746)

WHAT ARE SUBGROUP ANALYSES?

In randomized clinical trials, subgroup analyses evaluate the treatment effect (e.g., a hazard ratio [HR]) for a specific endpoint (e.g., overall survival) in subgroups of patients defined by baseline characteristics (e.g., age, gender, histology, and ethnicity). It is not recommended to base subgroups on postrandomization measures because the designation of patients to a subgroup may be affected by the study treatments.

Subgroup analyses are useful in endeavoring to obtain maximum information from a clinical trial by trying to identify subsets of patients that are more likely to benefit from the experimental treatment and conversely, by also detecting subsets of patients, which are at greater risk of being adversely affected. Subsequently, new hypotheses and trials can be generated from these findings. Ultimately, this may lead to changes in clinical practice. In addition, subgroup analyses can be useful in investigating whether overall treatment effects (e.g., increased efficacy or tolerability of the new treatment over the comparator) are consistent across subsets of patients. This is commonly referred to as "robustness checking." For these reasons, regulatory guidelines endorse appropriate subgroup analyses to be performed.[1–4]

WHAT ARE THE PROBLEMS WITH SUBGROUP ANALYSES?

There are two key statistical limitations of subgroup analyses. First, they are frequently underpowered. This is because the sample size of a clinical trial is calculated to evaluate the primary objective of the study with sufficient power in all randomized patients, not in a subset of patients. Hence, the interaction test to detect whether the treatment effect observed in one level of a subgroup (e.g., males) is significantly different to that observed in another level of the subgroup (e.g., females) is often underpowered. Consequently, subgroup analyses are prone to generating "false-negative" results.

The second major limitation of subgroup analyses is that they are particularly prone to multiplicity. Multiplicity is the inflated probability of getting a "false-positive" result, i.e., incorrectly concluding that there is a significant difference between treatment arms where one does not in fact exist, when several comparisons are performed. For example, when the primary objective of a trial is analyzed, this represents one comparison of the treatment arms. A 5% probability of obtaining a false-positive result is accepted as the null hypothesis is rejected if the p value is less than 0.05.

As more comparisons of the treatment arms are made, by performing multiple subgroup analyses of the primary endpoint, there is a greater chance of one or more of these comparisons generating a significant result by chance alone. For example, if 10 comparisons of the primary endpoint were done, there is a 40% chance of at

*Intercontinental Information Sciences, Eli Lilly and Company, Sydney, Australia; †Division of Oncology, Department of Medicine, Washington University School of Medicine; and ‡Alvin J Siteman Cancer Center at Washington University School of Medicine, St Louis, Missouri.

Disclosure: Helen Barraclough, MSc, is employed by Eli Lilly Australia and holds stock in Eli Lilly and Company.

Address for correspondence: Ramaswamy Govindan, MD, Division of Medical Oncology, Washington University School of Medicine, 660 S. Euclid, Box 8056, St Louis, MO 63110. E-mail: rgovinda@im.wustl.edu

ISSN: 1556-0864/10/0505-0741

Information that MUST be prespecified:

- The endpoint to be analyzed, e.g., overall survival.
- The baseline characteristic that defines the subgroup, e.g., gender.
- Statistical method used to test for an interaction between treatment and subgroup.

Additional information which it is good practice to also specify:

- Rationale for conducting the subgroup analysis, e.g., Biological hypothesis, previous result from another study.
- The levels of the baseline factor defining the subgroup, e.g., males and females.
- Expected direction of the treatment effect for each of the levels of the baseline factor defining the subgroup, e.g., experimental treatment is better than the comparator in females.

BOX 1. Information to document when prespecifying a subgroup analysis.

least one of these giving a false-positive result. Hence, a *p* value of less than 0.05 in a single comparison does not provide adequate evidence that there is a significant difference between treatment arms when multiple subgroup analyses are performed.

HOW NOT TO DO A SUBGROUP ANALYSIS

Subgroup analyses can sometimes be presented to "save" a failed study. This is when the primary objective of the trial was not met, but the new treatment was found to be significantly better than the comparator in a particular subset of patients. Many subgroups would have been analyzed to try to find the one (or a few) subsets(s) of patients in which the new treatment was significantly better than the comparator. This is sometimes described as "data dredging" or a "fishing trip." Misinterpretation of subgroup analyses can initiate future research based on unsubstantiated hypotheses and can even eventuate in suboptimal patient care.[5] These detrimental consequences are extremely costly but can easily be prevented by understanding the basic principles of subgroup analyses.

HOW TO CORRECTLY CONDUCT AND INTERPRET SUBGROUP ANALYSES

To conduct and interpret a subgroup analysis appropriately, it first needs to be established whether the subgroup analysis was prespecified. This is because the purpose of prespecified and unplanned subgroup analyses are distinct. Prespecified subgroup analyses are used for hypothesis testing. In contrast, unplanned (also called exploratory, retrospective, or posthoc) subgroup analyses are used for generating new hypotheses and for "robustness checking." It is imperative to understand that both can provide valuable information but for different reasons. Conclusive inferences and any subsequent changes in clinical practice can only be made from prespecified subgroup analyses. Hence, the remainder of this article will focus on how to appropriately perform and interpret prespecified subgroup analyses only. To overcome the two major statistical limitations of multiplicity and reduced power described above, the following five steps outline the best way to appropriately carry out, interpret, and report prespecified subgroup analyses: (i) prespecify the subgroup analysis in the protocol and/or the statistical analysis plan (SAP), (ii) use an interaction test, (iii) estimate the treatment effect for each level of the subgroup, (iv) validate results using confirmatory evidence, and (v) report results responsibly.

Prespecify the Subgroup Analysis in the Protocol and/or the SAP

Prespecified subgroup analyses are documented before any inspection of the data, whereas unplanned subgroup analyses are not. In most cases, prespecified subgroup analyses will be recorded in the protocol. However, they can also be detailed in the SAP before unblinding of the data or before first patient visit in open-label studies. Box 1 outlines the information that should be documented when prespecifying a subgroup analysis.

Prespecified subgroup analyses are regarded as more credible because they were planned before any examination of the data. This provides reassurance against "data dredging." However, both prespecified and unplanned subgroup analyses are prone to multiplicity, that is, the increased probability of a false-positive result because of testing multiple subgroups described above. Hence, simply prespecifying a subgroup analysis does not make it automatically valid: it must still be conducted, interpreted, and reported appropriately as outlined by the following steps.

Use an Interaction Test

Interaction tests are the most appropriate statistical method for conducting subgroup analyses. The concept of an interaction test can be illustrated with the following hypothetical example. In a randomized clinical trial, there are two treatments arms: treatment A (Tx A) and treatment B (Tx B), and the primary endpoint is overall survival. Gender is the baseline characteristic used to define the subgroup into two levels: males and females.

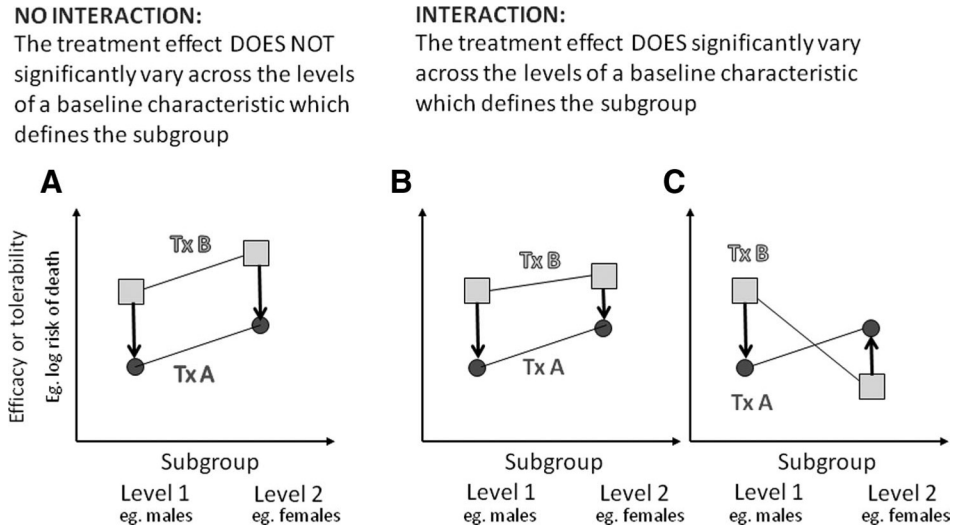

FIGURE 1. What is an interaction test? In this hypothetical example, there are two treatment (Tx) arms in the clinical trial: A and B. There are also two levels of the subgroup of patients defined by the baseline characteristic of gender: males and females. The regression lines linking the circles and squares represent the efficacy of treatment A and B, respectively, for overall survival (as estimated by the log hazard from the Cox Proportional Hazards Model). The log hazard estimates the log risk of death. Hence, the higher the regression line, the higher the risk of death. The treatment effect is illustrated by an arrow in each level of the subgroup (which in this example is the log HR[Tx A vs Tx B]). If the regression lines are parallel, there is no interaction between treatment and gender (*A*). Hence, the treatment effect in males is the same as in females. However, if the regression lines are not parallel (*B* and *C*), there is a statistically significant interaction between treatment and gender. Thus, the treatment effect in males is significantly different to that observed in females.

A significant interaction test shows that the treatment effect in males is not the same as in females (Figure 1). In the case of a nonsignificant interaction test, the treatment effect observed in males is not significantly different to the treatment effect observed in females. In this example, both males and females treated with treatment A had better overall survival than those patients treated with treatment B. This is shown in Figure 1*A* by the estimate for treatment A being lower than that for treatment B as the risk of death on treatment A is lower than on treatment B. The magnitude of the overall survival improvement observed with treatment A compared with treatment B was also the same in both males and females (as shown by the identical arrows in Figure 1*A*).

A significant interaction test shows that the treatment effect significantly varies across the levels of the subgroup. This can be described as either a "quantitative" or "qualitative" interaction (it may also be called heterogeneity). Figures 1*B* and 1*C* illustrate two scenarios where the interaction test was significant. In Figure 1*B*, both males and females assigned to treatment A experienced better overall survival than those assigned to treatment B. However, the size of the treatment effect was smaller in females than in males (as shown by the shorter arrow for females). This is an example of a "quantitative interaction." In Figure 1*C*, males had better overall survival when assigned to treatment A, but females experienced worse overall survival when assigned to treatment A (because their risk of death is higher). Hence, in this example, the direction of the treatment effect in males was opposite to that observed in females (as shown by the arrows pointing in different directions). This is an example of a "qualitative interaction."

An interaction test is usually carried out as part of a regression model. The type of regression model depends on the endpoint being analyzed. For "time-to-event" endpoints, such as overall survival and progression-free survival, a Cox Proportional Hazards model is used, whereas for binary endpoints, such as tumor response rate, a logistic regression model is used. The Cox Proportional Hazards Model is the standard method for analyzing time-to-event endpoints in clinical trials.[6] Therefore, in the case of this hypothetical example, the "treatment-by-gender" interaction test is carried out by using a Cox model containing:

- A treatment term (treatment A vs treatment B)
- A gender term (males vs females)
- A treatment-by-gender interaction term (males assigned treatment A vs all other patients)
- Plus any predefined prognostic factors based on baseline patient and disease characteristics (optional)

The interaction HR is a ratio of two HRs:

$$\frac{\text{HR (treatment A vs treatment B) for males}}{\text{HR (treatment A vs treatment B) for females}}$$

This can be alternatively written as:

$$\frac{\text{HR (males vs females) for patients assigned treatment A}}{\text{HR (males vs females) for patients assigned treatment B}}$$

The test for interaction has the null hypothesis that the interaction HR = 1, i.e., the treatment effect in males is the same as in females. The Cox proportional hazards model provides an estimate of the interaction HR and an associated *p* value.

If the *p* value for the interaction test is statistically significant, the null hypothesis can be rejected and a significant "treatment-by-gender interaction" can be claimed. Hence, the interaction HR differs significantly from 1. This means that the treatment effect observed in males is significantly different to the treatment effect observed in females. The size and direction of the treatment effect, i.e., HR (Tx A vs Tx B), can now be estimated for males and for females. If the interaction test result is nonsignificant, a differential treatment effect is not found, and thus, further analyses to test a predefined hypothesis are not recommended.

Estimate the Treatment Effect in Each Level of the Subgroup

An estimate of the treatment effect in males and in females can be obtained from either (i) the same Cox model described above or (ii) by removing the gender term and the "treatment-by-gender" interaction term and rerunning the model for males only and then separately for females.

Both approaches provide a HR (Tx A vs Tx B), 95% confidence intervals, and an associated *p* value for each level of the subgroup. These are often presented on a forest plot (Figure 2). From the estimated HRs in males and females, it can be determined whether the interaction is "quantitative" (Figure 1*B*) or "qualitative" (Figure 1*C*). If the interaction is "quantitative," the HRs would be in the same direction, e.g., less than 1, for both males and females. In contrast, if the interaction is "qualitative" then the HRs would be in opposite directions for each level of the subgroup, e.g., a HR(Tx A vs Tx B) <1 for males and a HR(Tx A vs Tx B) >1 for females.

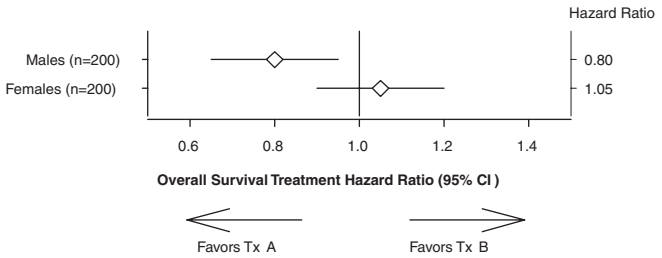

FIGURE 2. Forest plot. Forest plot are commonly used to graphically present subgroup analyses results. Above is a hypothetical result corresponding to the qualitative interaction example described in Figure 1C. The diamond represents the point estimate of the HR(Tx A vs Tx B) and the horizontal lines the 95% confidence intervals.

The associated *p* value of the HR in each level of the subgroup should be interpreted with caution. For example, suppose the associated *p* value = 0.001 in males and *p* = 0.08 in females. These *p* values give the probability of observing the estimated treatment difference or a more extreme one in each level of the subgroup by chance alone, given the null hypothesis that there really is no treatment difference is true. A common mistake is to claim that there is a differential treatment effect because the *p* value associated with the HR is statistically significant in males but nonsignificant in females. This is incorrect because only the interaction test *p* value determines whether the HR observed in males is significantly different to the HR observed in females. This is because the interaction test takes into account: (i) the prognosis of patients in different levels of the subgroup, e.g., females may have better overall survival than males regardless of the treatment they were assigned and (ii) the intergroup variability between males and females in addition to the intragroup variability.

Validate Subgroup Results Using Confirmatory Evidence

Validation of results is a fundamental scientific principle. To confirm a subgroup result from an individual clinical trial, presence of the subgroup effect in an independent study or meta-analysis is required. Additional, but less compelling types of confirmatory evidence that may be used to support the validity of a subgroup analysis result include a prespecified biologic rationale and the existence of the subgroup effect for related endpoints. It should be emphasized that until confirmatory evidence exists to validate a subgroup analysis result, it is hypothesis generating only and the treatment effect observed in all randomized patients is still regarded as the most appropriate estimate for patients in each level of the subgroup.

Report Results Responsibly

Subgroup results need to be reported responsibly for others to be able to interpret them appropriately. The results of the primary endpoint analysis in all randomized patients should be emphasized in abstract and conclusions. Furthermore, the prespecified subgroup analyses should be named, and the number of prespecified and unplanned subgroup analyses that were carried out should be clearly stated. The validity of a subgroup analysis result should also be discussed in context of current confirmatory evidence and the scientific literature.

SUMMARY

These concepts apply to any type of endpoint, such as categorical (e.g., responder or nonresponder), continuous (e.g., systolic blood pressure), or time to event data (e.g., overall survival). Box 2 summarizes the key points to aid clinicians to interpret subgroup analyses correctly.

Important concepts:

- Define subgroups based on baseline characteristics.
- Prespecified and unplanned subgroup analyses can both provide valuable information but for distinct purposes. Prespecified subgroup analyses are hypothesis testing, whereas unplanned subgroup analyses are hypothesis generating. Hence, they are interpreted differently.
- Only prespecified subgroup analyses can lead to changes in clinical practice.
- Unplanned subgroup analyses are valuable to generate new hypotheses, which can be tested by future research and to investigate the consistency of trial outcomes across different subsets of patients ("robustness checking"). Interpret unplanned subgroup analyses with caution and do not confuse these with prespecified subgroup analyses.
- Limit the total number of subgroup analyses carried out (whether planned or unplanned), by using available scientific rationale as a basis for any subgroup analysis.

Major points about conducting and interpreting prespecified subgroup analyses:

- Prespecify and justify a limited number subgroup analyses in the protocol if previous study findings or biological hypotheses are available.
- Use an interaction test to determine whether the treatment effect significantly varies across the levels of the subgroup.
- If the interaction test *p* value is statistically significant, proceed to interpret the size and direction of the estimated treatment effect for each level of the subgroup to determine whether there is "qualititative" or "quantative" interaction.
- If the interaction test *p* value is not statistically significant conclude that the treatment effect is NOT significantly different in one level of a subgroup (e.g., males) to another (e.g., females). Proceeding to estimate the magnitude and direction of the treatment effect in each level of the subgroup is not recommended, although "robustness checking" may still be carried out as deemed appropriate. If this is done, the aforementioned conclusion (of no interaction) still holds true regardless of how different the treatment effects may seem and how significant the *p* values associated to the hazard ratios are in each level of the subgroup.
- Use confirmatory evidence to validate prespecified subgroup results. Prespecified subgroup analyses are not automatically genuine or robust.
- The treatment effect observed in all randomized patients is regarded as the most appropriate estimate for each level of a subgroup if: (a) the interaction test is not statistically significant or (b) the interaction test is statistically significant, but insufficient confirmatory evidence is currently available to validate the subgroup result.
- Report results responsibly with sufficient information for others to make informed conclusions.

BOX 2. Key points of subgroup analyses in randomized clinical trials.

ACKNOWLEDGMENTS

The authors thank Lorinda Simms, Nicolas Scheuer, and Mauro Orlando for helpful discussions and critical reading of the article.

REFERENCES

1. International Conference on Harmonization (ICH) Topic E9. Statistical Principles for Clinical Trials, 1998.
2. Committee for Proprietary Medicinal Products (CPMP). Points to Consider on Multiplicity Issues in Clinical Trials (CPMP/EWP/908/99). London: EMEA, 2002.
3. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med* 2001;134:657–662.
4. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663–694.
5. Lagakos SW. The challenge of subgroup analyses—reporting without distorting. *N Engl J Med* 2006;354:1667–1669.
6. Cox DR. Regression models and life-tables. *J Royal Stat Soc Ser B* 1972;34:187–220.